SCinet: Testbed for High-Performance Networked Applications

Once each year, leading experts in ultra-high-performance networking and computing spend a week building the world's fastest network and running applications that stress its capabilities.

William T.C. Kramer
Lawrence Berkeley National Laboratory

The rapidly increasing availability of high-performance networking is arguably computing's most significant contribution to society in the past 20 years. Despite network availability, it is difficult to evaluate how well high-performance applications use its complex, multilevel interconnections in the real world.

To better understand this interaction, leading experts and organizations in ultra-high-performance networking and computing come together once a year to build what is likely the world's most intense, diverse, high-performance network. Almost overnight, the SCinet testbed comes online to showcase state-of-the-art network technology at the IEEE/ACM-sponsored supercomputing conference known as SCxy. The testbed is in place for four days, during which time teams of application developers demonstrate bandwidth-intensive applications that stress the network's capabilities. At the end of the demonstrations, SCinet is dismantled, and work begins on the next year's design.

SCinet runs hundreds of bandwidth-intensive applications each year, but SC2000 and SC2001 featured a special "Network Bandwidth Challenge" for applications to try making full use of SCinet's wide-area capabilities. In the words of one network engineer, the Bandwidth Challenges asked application developers to "burn down the world's fastest network." The SCinet team selected 10 of the most data-intensive applications in 2000 and 12 applications in 2001 for formal evaluation, and Qwest provided prizes for the winners.

SCINET ARCHITECTURE

SCinet is built from loaned networking equipment and services—more than $25 million each year for 2000 and 2001, as well as the donated effort of more than 100 leading network engineers. The SC2000 network was created and fully operational in just over five days, and consisted of 82 miles of fiber optic cable installed in less than 51 hours. SC2001 took similar amounts of fiber and time.

At the SC2000 conference in Dallas, the local area network (LAN) within the SCinet testbed area had a peak capacity of 130 gigabits per second. The network connected to all the major national scientific networks and supercomputer centers, and maintained a total wide-area network (WAN) bandwidth of almost 9 Gbps. At SC2001 in Denver, the LAN capacity increased by 28 percent—to 194 Gbps—and the WAN capacity almost doubled—to 15.7 Gbps.

SCinet is an Internet service provider in its own right. Its network design is intentionally complex to explore issues typically encountered in real-world networking, such as interoperating between different network domains and across different routers and technologies. The overall network design for SC2000 and SC2001 consisted of four network levels that were all interconnected, but could operate independently of each other:

- a commodity Internet network for conference business functions,
- an 11-Mbps wireless network spanning the entire conference area,

0018-9162/02/$17.00 © 2002 IEEE

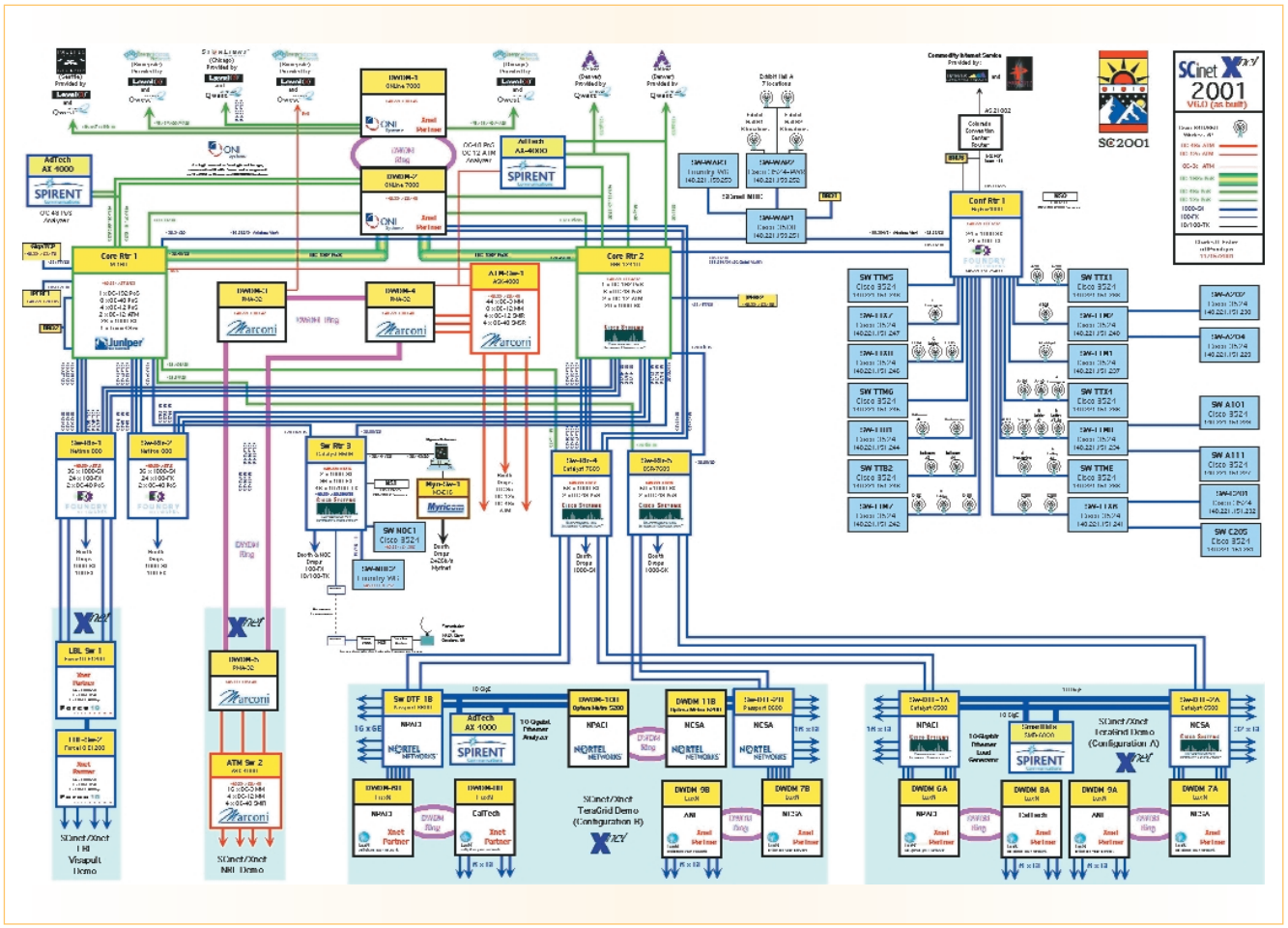

Figure 1. The SCinet 2001 network architecture. The commodity Internet network (top right) includes 11-Mbps wireless network. The SCinet production network (top left) connected exhibitors to the major external wide-area networks through two core routers and OC-192 connections. The Xnet experimental network (bottom) demonstrated 10-Gigabit Ethernet connectivity separate from the commodity and production networks. Image provided by Charles Fisher of Oak Ridge National Laboratory, 2001.

- the SCinet production network, which connected the exhibitors to the major external WANs, and
- the Xnet experimental network.

Figure 1 shows the overall network design for SC2001.

Table 1 shows the high-bandwidth connectivity within the network for each year. The changes from 2000 to 2001 indicate shifts in the technology away from asynchronous transfer mode (ATM), Packet over Sonet (PoS), and Fast Ethernet to Gigabit Ethernet, wireless, and DWDM.

Table 2 shows SCinet's external wide-area bandwidth, which increased by 66 percent from SC2000 to SC2001. In the external WAN architecture for SCinet 2000, most of the peering traffic—and indeed nine of the 12 Bandwidth Challenge applications—used the High Speed Connectivity Consortium (HSCC) network to route to the National Transparent Optical Network (NTON) and other networks. HSCC actually routed traffic from SCinet over the Qwest backbone network.

To avoid affecting the backbone traffic for Qwest's many paying clients, SCinet agreed to limit traffic over the HSCC link to 1.5 Gbps of real traffic. Automatically limiting traffic from one application to another in such a complex environment is not yet technically feasible; therefore, the SCinet and HSCC staff had to implement this threshold by monitoring demonstration traffic and manually throttling the applications that used HSCC. This turned out to be the major performance limitation for some applications. For example, SCinet 2001 had no such limitation, and Visapult—which won awards in both years—proved capable of using more than twice the bandwidth than it had in 2000.

BANDWIDTH CHALLENGE APPLICATIONS

The conference demonstrates several hundred scientific and technical applications each year, about half of which rely on high-performance networking. Of the 22 applications featured in the Network Bandwidth Challenge competitions at SC2000 and SC2001, the three winners for each year fall into

four broad application categories: remote visualization, quality of service, high-performance data transfer, and real-time collaboration. (For a summary description of all 22 competitors, see http://www.nersc.gov/~kramer/SCinet.)

Remote visualization

Researchers have used advanced networking for many years to visualize data remotely from the systems that compute and store it. Despite the remarkable speedup in desktop systems, the data sets that visualization tools examine grow even faster. These tools must take full advantage of the network to reach the data needed and then visualize it at the scientist's location. Transmitting a single file in multiple streams via parallel data transfer is one way to achieve this performance, but it challenges the application to keep track of all the streams and to reassemble them correctly at the source.

SC2000 Fattest and Fastest. The SC2000 application that sustained the fastest wide-area bandwidth and transferred the most data in a fixed amount of time was Visapult, a prototype distributed application for the remote visualization of terabyte data sets.[1] Visapult employs parallel components that communicate with one another to achieve the high data throughput needed for interactive visual analysis. A key service layer in the SC2000 demonstration was the Distributed Parallel Storage System, which operates at the speed of locally attached storage regardless of the actual location of the data on the WAN.[2]

In the SC2000 demonstration, the Visapult team used

- a data server running DPSS at the Lawrence Berkeley National Laboratory,
- an eight-CPU SGI Origin computer provided by the Accelerated Strategic Computing Initiative (ASCI), which ran Visapult on the testbed floor in Dallas to visualize an 80-Gbyte data set remotely; and
- dpss_get, a high-speed parallel file transfer application running on an eight-node Linux cluster provided by Argonne National Laboratory.

The Visapult team received the Fattest and Fastest award when the application recorded a peak performance level of 1.48 Gbps across the 2.5-Gbps wide-area link they used to access resources throughout the US. As noted earlier, however, SCinet artificially limited the bandwidth across this link to a 1.5-Gbps maximum.

Table 1. SCinet local area network connections by type.

Type of connection	Number at SC2000	Number at SC2001
10-Gigabit Ethernet	1	3
OC-192 PoS		2
DWDM Rings		7
OC-48c ATM	6	
OC-48 PoS	5	1
OC-12c ATM	13	2
OC-12 PoS	2	
OC-3c ATM	7	2
1-Gbps Ethernet	72	64
Fast Ethernet	79	53
802.11b Wireless Access Points	27	39

Table 2. Total external wide-area network connection types, partners, and bandwidths for SC2000 and SC2001.

Network Type	SC2000 Network Partners	SC2000 Maximum Bandwidth	SC2001 Network Partners	SC2001 Maximum Bandwidth
OC-48	PoS Abilene/ Internet2, HSCC	2*2.5 Gbps	Abliene/Internet2, ESnet, Starlight, Pacific Gigapop	6*2.5 Gbps
OC-48c ATM	ATDnet	2.5 Gbps		
OC-12 ATM	Esnet, vBNS	2*655 Mbps	ESnet	655 Mbps
OC-12 PoS	vBNS	655 Mbps		
ATM	Commodity	12 Mbps	Commodity	100 Mbps
	TOTAL	9.477 Gbps		15.755 Gbps

Overall, the team transferred 262 Gbytes of data in 60 minutes from the Berkeley Lab to the Dallas show floor. Their 60-minute average throughput of 582 Mbps included startup time, application tuning, and adapting for problems.

SC2001 Fattest and Fastest. Visapult won the Fattest and Fastest award again in 2001, this time in combination with a data service to study complex astrophysical phenomena. The Cactus Computational Toolkit (http://www.cactuscode.org) established a computational grid of software components, including parallel I/O, visualization, and the computational tools necessary to study strong dynamical gravitational fields.[3]

Besides staff from Berkeley, the SC2001 demonstration included collaborators in Illinois and Germany. It used a live feed of simulation data from the Cactus code to visualize colliding black holes, computed in real time at supercomputing centers in Berkeley and Champaign, Illinois. The Cactus code is among the most complex and resource-demanding calculations in the world. Developed by the Albert Einstein Institute's Numerical Relativity Group in Potsdam, Germany, the code also won a Gordon Bell Prize for high-performance computing at the conference.

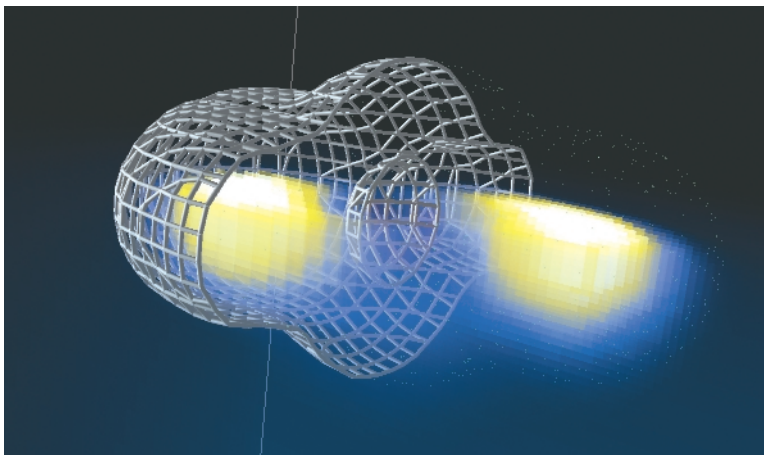

Figure 2. Computer simulation and visualization of gravitational waves from inspiraling coalescing black holes. Image provided by John Shalf and Wes Bethel of Lawrence Berkeley National Laboratory.

The demonstration achieved a sustained performance level of 3.3 Gbps. The computational resources included a 3,328-processor IBM SP2 at the National Energy Research Scientific Computing (NERSC) Center in Berkeley, one of the largest unclassified supercomputers in the world, with a peak performance of 5 teraflops per second, as well as a 128-processor SGI Origin 2000 located at the National Center for Supercomputing Applications. The demonstration received and processed data on the SC2001 show floor through a 32-way parallel Sun StarFire system and eight-node Linux cluster, fed by a 10-Gigabit Ethernet switched from Force10 Networks.

Figure 2 shows a sample frame from the visualizations with a component of the gravitational radiation emanating from the merger. Researchers see black-hole coalescence as a promising source of detectable gravitational waves, which in turn are expected to either support or refute Einstein's 80-year-old General Theory of Relativity. Simulations of these phenomena illuminate the signal processing and detection techniques required to maximize the efficient use of gravitational wave detectors, such as the Laser Interferometer Gravitational Wave Observatory, when these very expensive experimental apparatuses come online.

Quality of service

In general, digital networks transmit data in packets that receive the same priority at each point along a route. It is possible, however, to give priority treatment to some packets. QoS attempts to use this capability to give an application consistent, predictable data delivery at an agreed-upon service level. QoS is an important network research and development area that requires cooperation at all network layers to minimize delivery delay and variation.

SC2000 Most Captivating and Best Tuned. A QoS-enabled audio teleportation demonstration won the SC2000 Most Captivating and Best Tuned award. The project, put together by Chris Chafe of Stanford University's Center for Computer Research in Music and Acoustics, showed the benefits of QoS when applied to the real-time transmission of interactive CD-quality audio across the network path between Stanford in Palo Alto, California, and the conference in Texas (http://apps.internet2.edu/html/qos-enabled-audio-demo.html).

The demonstration used music because human hearing is very sensitive to dropped or delayed information in music, particularly when played on fine instruments. Most Internet music systems today rely on buffering a stream of data at the listening point to ensure continuous, quality sound. Instead, this application relied on QoS routing to ensure that packets associated with the application received priority as they traveled across national networks.

A musician played a string instrument at the testbed floor, but rather than picking up the sound through an amplifier, the demonstration captured the live sound electronically, converted it to digital format, then transmitted it to a stairwell at Stanford, where it was played. The stairwell acted as an echo chamber, giving the music an additional sound quality. In the stairwell, microphones picked up the music, redigitized it, and transmitted it back to Dallas, where it was played to the audience. The audience in Dallas heard the stairwell-amplified sound, which appeared to come directly from the performance in front of them.

Network QoS for this demonstration consisted of marking application traffic for expedited forwarding (EF), shaping and policing it at the network edge, a design that reflects the architecture of the Internet2 QBone Premium Service. The audio streams traversed segments that tested preferential service to EF-marked traffic before being sent over the Stanford CalREN2 research network connection and the backbone.

The demonstration showed effective protection of the application traffic through heavy congestion artificially induced near the network edge. For comparison purposes, the demonstration dynamically enabled and disabled the QoS configuration to show the differences in the original sound and the augmented sound without QoS protection.

High-performance data transfer

A new approach to the complex, high-performance infrastructure that large-scale science applications rely on—the Grid—is making a tremendous impact on scientific computing.[4] A set of software, as well as a concept of use, the Grid has evolved rapidly over the past four years. It stems from work done on the Globus Toolkit (http://www.globus.org) and aims to simplify access and

coordinated use of large-scale resources such as supercomputers, terascale data storage systems, and experiment devices. The Grid has the potential to unleash a new generation of software applications that assume flexible access to distributed resources.

SC2000 Hottest Infrastructure. Grid applications rely on very high performance underlying networks to access resources and move data to the systems doing a computation. Several Bandwidth Challenge demonstrations used Grid software, including the SC2000 Hottest Infrastructure award winner—an application in climate modeling research.[5] This project demonstrated an infrastructure for secure data transfer as well as replication of large-scale climate modeling data sets.

The data sets for climate modeling applications consist of many files, ranging to many gigabytes in size, that are often duplicated at various locations. When a researcher requests a particular view of the data, Grid software transfers relevant files from the data replica that offers the best performance.

This SC2000 application included several components. First, users specify the desired data's high-level characteristics—for example, precipitation amounts for a certain time period and region. Then, Grid infrastructure software maps a metadata infrastructure between these high-level attributes and file names. Next, a replication management infrastructure finds the physical locations of all the files.

The user selects among these locations by consulting performance and information services such as the Network Weather Service and the Globus Toolkit's Monitoring and Directory Service to predict relative performance of transfers from each location. By selecting a particular physical replica, the user initiates secure, high-performance data transfer between the source and destination sites. Finally, the application presents the desired data graphically to the user.

This SC2000 project resulted from the collaboration of three groups. Researchers at Lawrence Berkeley National Laboratory created a request manager that calls low-level services and selects among replicas. Scientists at Lawrence Livermore National Laboratory provided the user interface and visualization output for the application as well the metadata service that maps between high-level attributes and files. Finally, the Globus project team at the University of Southern California's Information Sciences Institute and the Argonne National Laboratory provided basic Grid services, including replica management, information services, and secure, efficient data transfer.

Real-time collaboration

Two applications of real-time collaboration picked up the remaining awards at SC2001.

SC2001 Most Courageous and Creative. James Oliverio and Andy Quay led "Dancing Beyond Boundaries," an intercontinental collaborative performance organized by the University of Florida's Digital Worlds Institute. The performance featured dancers in Denver, Minneapolis, and Gainesville, Florida, accompanied by musicians in Brazil, and won the SC2001 Most Courageous and Creative award (http://www.digitalworlds.ufl.edu/sc2001/).

The project explored artistic collaboration among internationally distributed dancers, musicians, graphic artists, videographers, and choreographers who created, rehearsed, and performed a new work using multiple network-conferencing nodes and a select number of high-quality video and audio streams.

SC2001 Best Network-Enabled Application. A different style of real-time application controlled a high-energy electron microscope in San Diego from the conference floor. A team led by Tom Hutton and Mark Ellisman from the San Diego Supercomputer Center (SDSC) and the University of California at San Diego's National Center for Microscopy and Imaging Research (http://ncmir.ucsd.edu) operated the instrument through a live video stream on an end-to-end IPv6 connection through SCinet to San Diego. The team does similar work with collaborators in Japan who view live video and control the microscope with advanced client-side Java applets.

The demonstration was the first use of SDSC's Storage Resource Broker (http://www.npaci.edu/dice/srb/) as a tool for moving large amounts of data over IPv6. The demonstrations highlighted interactive, collaborative Web portal access to key elements of the Grid infrastructure, including instrumentation, computation, and database elements, as well as collaborative tools.

Basic Grid services include replica management, information services, and secure, efficient data transfer.

NETWORK USAGE

SCinet uses some of the most advanced monitoring tools to measure network activity. High-bandwidth applications offer a special case for demonstrating these tools. Further, the tools overlap, making it possible to validate the accuracy of reported numbers—something that is rarely possible in the general environment. This gives vendors valuable feedback on their tools, whether beta-level

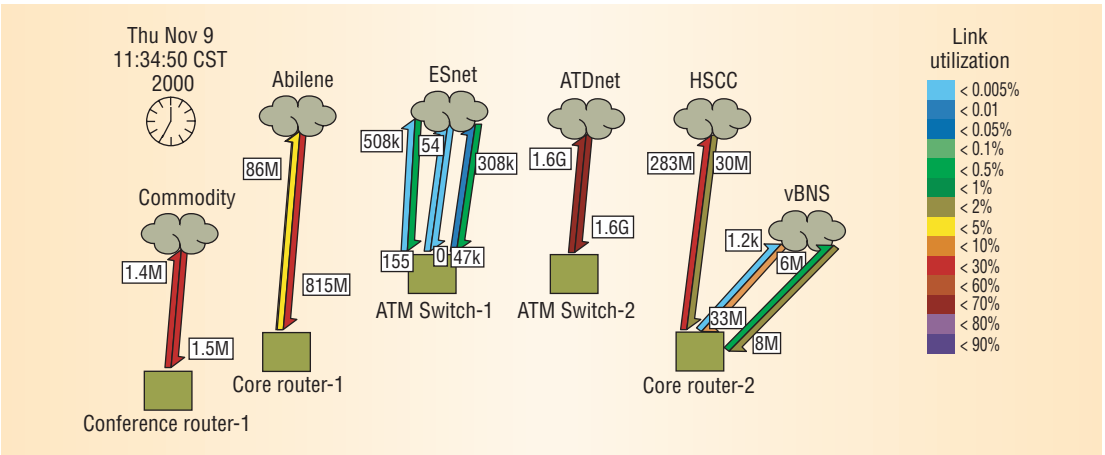

Figure 3. SCinet-2000 WAN weather map. Aggregate bandwidth use totals almost 4.4 Gbps out of a total capability of 9.3 Gbps. Monitoring software and data courtesy of the University of Florida's Greg Goddard and Internet 2's Matthew J. Zekauskas.

deployments of commercial systems or university projects.

Individual bandwidth measurements showed that at least one SC2000 application achieved over 3.2 Gbps on a sustained basis, transferring HDTV data streams with real-time control of images between Dallas and Washington, DC. The Visapult application transferred 1.56 Gbps in SC2000 on a 5-second sample rate, and 1.76 Gbps on a 0.1-second sample. At SC2001, the Visapult-Cactus application achieved a sustained rate of 3.3 Gbps.

SCinet used SNMP (simple network management protocol) to measure the 5-second sample by polling the routers involved. The 0.1 sample rate came from Spirent/Adtech measurement devices that directly monitor the application by physically tapping the links and then associating sockets and IP addresses for traffic analysis.

The three network performance measures—one within the application, one monitoring packets associated with the distributed parts of the application, and another monitoring the router—all showed agreement in the evaluation and measurement.

Total bandwidth

SCinet experiments with how much bandwidth the entire network can support in and out of the conference at the same time. Figure 3 shows a snapshot "weather map" of external network usage from Friday, 9 November 2000. Several bandwidth-intensive applications were running at the time, though not Visapult or other HSCC-bound applications. Even so, this measurement totaled almost 4.4 Gbps, or 51 percent of the theoretical maximum. The high-water mark for bandwidth usage was observed at 4.9 Gbps out of the maximum 8.5 Gbps—or a sustained 58 percent utilization over a 15-minute interval.

At SC2001, the highest rate observed was 5.9 Gbps; though higher than SC2000, this is only 37 percent of the peak. SC2001 had more connections to networks than SC2000, but the maximum OC-48 rate (655 Mbps) per connection remained the same. As a result, the aggregate bandwidth used increased in total, but the percent of the total used was lower in 2001 because the speed limit per link did not increase. Although it is clear that scientific applications can make effective use of any level of bandwidth available, routing data over multiple WANs remains a challenging problem that only a few applications can manage.

Security monitoring

Networks must also monitor and at times enforce computer security. Currently, the most common method places resources behind firewalls that filter all network traffic and limit access. Unfortunately, this method limits performance as well as functionality. At this time, no commercial firewalls can accommodate the data rates and functions SCinet uses.

Instead of a firewall, SCinet uses the Bro intrusion detection system developed by Vern Paxson at Lawrence Berkeley National Laboratory.[6] Rather than having all network packets flow through a firewall, Bro sits alongside the network, monitoring IP packets as they travel over the media and watching for suspicious traffic patterns. Its normal operating mode scans application-layer traffic, which minimizes the impact on network performance.

Bro includes powerful features to analyze scanned packet data. When Bro detects a suspicious pattern, such as a systematic scan of all IP addresses by an external source, it can automatically take actions to limit or block the traffic, or it can inform security experts that something should be examined in greater detail. Bro also keeps extensive logs of network behavior for later detailed examination. By maintaining a long-term view of traffic, Bro maps patterns rather than just predefined connections or events, giving it a distinct advantage over firewall technology.[7]

Bro required some modifications to scan the high-speed links used at SC2001 without affecting application performance. For links greater than 1 gigabit, Eli Dart and Rob Jaeger—network engineers at LBNL/NERSC and Juniper Networks, respectively—collaborated to have the Juniper router pass

filtered network traffic through a "side door" for analysis, rather than tapping the physical media. The result showed that a single Bro system attached through Gigabit Ethernet could monitor three OC-48 links if the router filtered the traffic at full network speed. This successful proof of concept will be carried further at SC2002.

Xnet

SCinet's commodity, wireless, and production levels must provide stable service. Vendors and researchers sometimes hesitate to showcase bleeding-edge hardware in a testbed network that applications use aggressively. Thus, SCinet incorporates an experimental network called Xnet, providing a context to demonstrate network equipment or capabilities that typically do not exist outside the development lab.

In 2000, the major Xnet project was one of the first public 10-Gigabit Ethernet demonstrations. It consisted of a point-to-point network arranged between the two show areas, using Cisco's pre-testbed 10-Gigabit Ethernet blades for their 6500 series switching routers with parallel optical interfaces. These interfaces short-circuit the full serialization process by intercepting the four parallel 10-Gbps interface streams and running them out directly as parallel data streams on optical ribbon cable.

Working with the ribbon cable was extremely difficult. Because four data streams needed to work, SCinet actually installed six separate spools of the cable. The demonstration showed a 20-CPU storage cluster in the SGI area feeding data through a pair of 10-Gigabit Ethernet cards to a 20-processor computing cluster in the ASCI area. Interfaces for each cluster consisted of 20 separate Gigabit Ethernet links.

In 2001, SCinet deployed three separate 10-Gigabit Ethernets that used Force10, Nortel, and Cisco switches. Many applications used these networks, including several Bandwidth Challengers. Additionally, SCinet-2001 deployed an Adtech network monitor on two of the 10-Gigabit Ethernet links to monitor traffic.

SCinet-2001 was also the first time the testbed used DWDM extensively, deploying seven rings. DWDM is optical technology that combines and sends different light wavelengths simultaneously on the same fiber, allowing one fiber to carry independent network connections along its path. DWDM is one reason people talk of bandwidth becoming virtually unlimited. It is already being used in several national networks, including NSF's Distributed Terascale Facility.

SCGlobal

SC2001 introduced SCGlobal, which used the SCinet infrastructure, along with Access Grid technology (http://www.accessgrid.org),[8] to link the core SC2001 activities with dozens of SCGlobal constellation sites. The constellation sites were distributed throughout North and South America, the Pacific Rim, and Europe, with one site at the South Pole.

SCGlobal aimed to provide a multinational and multicultural meeting place. It consisted of Access Grid nodes that used digitized video and audio to link presentations and sessions between the main conference and the constellation sites. The interaction was two-way with presentations originating from the remote sites as well as the conference.

The Imagine Team at Lawrence Berkeley National Laboratory created one of the more interesting Access Grid sites: a mobile robot that wandered through the conference exhibit hall. Shown in Figure 4, RAGE (Remote Access Grid Entity) featured the two-way video and audio characteristics of all Access Grid sites but used SCinet's 11-Mbps wireless network to remain constantly connected as it moved around the conference. One technical challenge was to scale Access Grid functionality and performance to the slower, more variable performance of a wireless network.

DWDM optical technology increases bandwidth by allowing one fiber to carry independent network connections along its path.

BANDWIDTH ISSUES

The Bandwidth Challenge showed that applications can be designed to make effective use of almost any amount of bandwidth conceivable today. Despite "the last mile" issues, which are at least as much regulatory as technical, it is feasible to connect almost any location at arbitrarily high speed.

While advanced applications and new infrastructure such as computational and data grids increase bandwidth requirements, new technologies like DWDM increase the amount of bandwidth available. Moreover, raw untapped bandwidth already exists in the US that can meet the needs of these applications—albeit for a price. So what limits the capabilities of large-scale applications to use the bandwidth in the nation's fiber infrastructure?

Moore's law

The first factor is the interface between the wide-area and local-area networks. Routers, firewalls, and other equipment at the WAN-LAN interface are computers, even if somewhat specialized. Moore's law limits their system performance. Yet, network bandwidth capacity on both sides of this

Figure 4. RAGE: Remote Access Grid Entity. A team from Lawrence Berkeley National Laboratory created a mobile Access Grid robot that demonstrated telepresence and remote control via a wireless Ethernet link for the SCGlobal showcase at SC2001.

interface increases at double or triple the rate of Moore's law.

While network bandwidth must increase in total to support more data from more connections and applications, any given end-to-end application must get data out of the system it is running on, through its local network, across the WAN-LAN interface onto the wide area, then off the wide area into the destination local network, and finally into the destination system. Data passes through at least two WAN-LAN interfaces, sometimes many more.

In addition, WAN-LAN interfaces include technologies such as network and computer security monitoring systems. Moore's law also limits the systems that implement these functions.

There are other performance challenges to getting data into and out of the end-point computers. Most of the key network functions are limited by the number of packets a system can process at a time, which in turn is often related to the number of interrupts a processor can handle. Processors use many levels of hardware enhancement to increase processor speed. Unfortunately, these enhancements also make processing individual packets more costly. They also make bridging the gap between internal and external networks more difficult.

Other end-point issues include the relatively slow I/O port speeds compared to processor speed. Further, most systems go through many layers of software to process network traffic. Finally, to fill up high-bandwidth networks, systems must hold larger and larger numbers of outstanding packets in memory buffers—now approaching Gbytes of data.

While new methods such as jumbo frames, interrupt coalescence, and specialized hardware interfaces address the challenges, the performance gap is still widening in most situations.

Network complexity

To make up for lost speed in the LAN-WAN interface, network applications parallelize their traffic, breaking it into multiple streams. Unfortunately, this runs into a second factor limiting network usage: inordinate network complexity and the corresponding management difficulties. The implementation of SCinet 2000 and its applications took more than 200 professionals and 11 work-years—not counting the efforts of vendors and service providers to provision the technology and help manage the national network, nor the effort involved in creating the applications. SC2001 took at least as much effort.

Networking and the Internet's "networks of networks" have become so complex that even the best network engineers cannot manage the entire network manually—not to mention the relative scarcity of network engineers overall. Sufficient tools do not yet exist to automate network functions, particularly across network administration boundaries. Consequently, to avoid disrupting service to its own users and others, each provider is conservative in its network design and use. Thus, most networks are planned so the actually traffic is a relatively small amount of the theoretical maximum. This means that the perceived amount of bandwidth is not really usable until we can manage it more effectively.

SCinet makes a major contribution to the networking field by bringing the resources and expertise together once a year to demonstrate what is possible. Without the investment in cross-network diagnostic and management software, it will become increasingly difficult to make multiple WANs work well together, and we will lose more efficiency in network capacity.

TCP/IP stress

Finally, there are fundamental protocol issues when operating TCP/IP at high rates. High-performance applications are moving to protocols that perform better but are not as "well behaved" and require the application to ensure reliable delivery. Such methods allow an application to send many packets into the network without the flow control associated with TCP/IP. These protocols, however,

can conflict and deny service to TCP packets, which could in turn affect the traffic generated by many common uses of the network. As network speed increases, TCP and IP will likely have to evolve to use bandwidth efficiently and to increase network bandwidth performance commensurately with computer and network hardware performance.

Despite the challenges of time, complexity, and evolving advanced technology, a dedicated group of experts implement the world's most powerful network and run real-world applications that stress its capabilities each year. This annual project in ultra-scale networking and computing is an encouraging indicator of the growth in this technology. The 22 Bandwidth Challenge applications at SC2000 and SC2001 not only used SCinet bandwidth and advanced features efficiently but also showed the need for increased network speed and function to unleash important applications to carry out scientific and collaborative missions in upcoming years. ■

Acknowledgments

Hundreds of people contribute to SCinet and the Bandwidth Challenges reported in this article. They are supported by many organizations. It is not possible to mention all by name, but special thanks must go to the following organizations and individuals (in alphabetical order) for their help on both the network and this article: Steve Corbato, Eli Dart, Paul Daspit, DOE Office of Science, Hal Edwards, ESnet, Chuck Fisher, Greg Goddard, Ian Foster, Internet2, Wesley Kaplow, Steve Lau, Jeff Mauth, Debbie Montano, Bill Nickless, Qwest, Jim Rogers, Martin Swany, John Shalf, Tim Toole, and Bill Wing, and of course the IEEE and the ACM for sponsoring the SC conference series.

My work is supported by the US Department of Energy's Office of Science, Division of Mathematical, Information, and Computational Sciences, under contract number DE-AC03-76SF0098.

References

1. W. Bethel et al., "Using High-Speed WANs and Network Data Caches to Enable Remote and Distributed Visualization," *Proc. Supercomputing 2000* (SC2000), ACM Press, New York, 2000; http://www.supercomp.org/sc2000/proceedings/techpapr/#10.
2. B. Tierney et al, "A Network-Aware Distributed Storage Cache for Data Intensive Environments," *Proc. IEEE High-Performance Distributed Computing Conference* (HPDC-8), IEEE CS Press, Los Alamitos, Calif., 1999; http://computer.org/proceedings/hpdc/0287/0287toc.htm.
3. G. Allen et al., "Supporting Efficient Execution in Heterogeneous Distributed Computing Environments with Cactus and Globus," *Proc. Supercomputing 2001* (SC2001), ACM Press, New York, 2001; http://www. sc2001.org/papers/pap.pap301.pdf.
4. I. Foster and C. Kesselman, eds., *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann, San Francisco, 1999.
5. W. Allcock et al., "High-Performance Remote Access to Climate Simulation Data: A Challenge Problem for Data Grid Technologies," *Proc. Supercomputing 2001* (SC2001), ACM Press, New York, 2001; http://www.sc2001.org/papers/pap.pap283.pdf
6. V. Paxson, "Bro: A System for Detecting Network Intruders in Real-Time," *Computer Networks*, vol. 31, nos. 23-24, 1999, pp. 2435-2463.
7. Y. Zhang and V. Paxson, "Detecting Stepping Stones", *Proc. 9th Usenix Security Symp*., Usenix Assn., Berkeley, Calif, 2000; http://www.icir.org/vern/papers/stepping/index.html.
7. L. Childers, "Access Grid: Immersive Group-to-Group Collaborative Visualization," *Proc. 4th Int'l Immersive Projection Technology Workshop*, 2000; http://www.ipd.anl.gov/anlpubs/2000/07/36282.pdf.

***William T.C. Kramer** is deputy director of the National Energy Research Scientific Computing Center and head of high-performance computing at Lawrence Berkeley National Laboratory. He also teaches in the Computer Science Department at the University of California, Berkeley. His professional interests include high-performance computing and networking, security, and system evaluation. He received an MS in computer science from Purdue and an ME in electrical engineering from the University of Delaware. Contact him at kramer@ nersc.gov.*